

cmenet: a new method for bi-level variable selection of conditional main effects

Simon Mak*, C. F. Jeff Wu^{*†}

January 20, 2017

Abstract

This paper presents a novel method for selecting main effects and a set of reparametrized predictors called conditional main effects (CMEs), which capture the conditional effect of a factor at a fixed level of another factor. CMEs represent highly interpretable phenomena for a wide range of applications in engineering, social sciences and genomics. The challenge in model selection lies in the grouped collinearity structure of CMEs, which can cause poor selection and prediction performance for existing methods. We propose a new method called **cmenet**, which employs coordinate descent and two principles called CME coupling and reduction to efficiently perform model selection. Simulation studies demonstrate the improved performance of **cmenet** over existing selection methods, such as the LASSO and SparseNet. Applied to a gene association study on fly wing shape, **cmenet** not only provides improved predictive performance over existing techniques, but also reveals important insight on gene activation behavior. Efficient implementations of our algorithms are available in the R package CMENET in CRAN.

Keywords: Coordinate descent, gene association, majorization-minimization, multi-collinearity, variable selection.

*School of Industrial and Systems Engineering, Georgia Institute of Technology

[†]Corresponding author

1 Introduction

This paper proposes a new method for selecting *main effects* (MEs) and a set of reparametrized predictors called *conditional main effects* (CMEs) from observational data whose factors have two levels.¹ A CME can be described as follows. Let A and B denote two binary factors with levels $-$ and $+$. The CME $A|B+$ is then defined as the effect A when effect B is at the $+$ level, and 0 when B is at the $-$ level. In words, such an effect quantifies the influence of A *only* when B is at the level $+$. The CME $A|B-$ can be defined analogously.

The appeal for CMEs as basis vectors for variable selection comes from its interpretability in a wide range of applications, including genomics and the social sciences. For example, in gene association studies, where the goal is to identify important genetic contributions for a trait or disease, the CME $A|B+$ quantifies the significance of gene A *only when* gene B is present. Here, the selection of CMEs can serve as an effective tool for investigating the activation behavior of gene-gene interactions, namely, which genes are *conditionally* active, and which are important in *activating* other genes. For single nucleotide polymorphisms (SNPs), these conditional gene interactions are known to be highly influential for certain traits (see, e.g., Chari and Dworkin, 2013; Leung et al., 2014). CMEs also arise naturally in many engineering applications. For example, in an injection molding experiment with two settings for mold temperature A and holding pressure B (pg. 352 of Montgomery, 2008), the CME $A|B+$ measures the effectiveness of mold temperature *only at* a high level of holding pressure. This conditional effect may be a result of material properties for the molding liquid, and the discovery of such effects can provide valuable insight on the injection process.

The idea of CMEs was first introduced in Wu (2015) as a way to disentangle effects which are aliased (i.e., perfectly correlated) in a *designed* experiment. Ever since the

¹While, for brevity, most of the methodology presented in this paper considers only two-level factors, the method is easily extendible to datasets with *both* two-level and continuous factors. We discuss such an extension in Section 3.4.

pioneering work of Finney (1945), it has been widely accepted in the design community that aliased effects in a regular, two-level design cannot be “de-aliased” without adding more experimental runs. Such a belief was shown to be false in Wu (2015), where the author employed a reparametrization of the aliased effects into CMEs to allow for selection of the correlated conditional effects. A variable selection method for designed experiments is further developed in Su and Wu (2017), which makes use of the natural groupings of CMEs into so-called twin, sibling and family effects. In this paper, we generalize this selection framework to *observational data*, by exploiting the implicit structure of CMEs to form new effect groups and to motivate a novel penalized selection criterion.

For penalized variable selection methods, the usual procedure for two-level factors is to first normalize each factor to zero mean and unit variance (Tibshirani, 1997). Treating the rescaled predictors as continuous variables, standard variable selection techniques using the l_1 -penalty in LASSO (Tibshirani, 1997) or non-convex penalties (e.g., Frank and Friedman, 1993; Fan and Li, 2001; Zhang, 2010) can then be used to pick out significant effects. For the problem at hand, however, such methods are inappropriate, because they do not account for the implicit group structure present in CMEs. Grouped selection techniques, such as the group LASSO (Yuan and Lin, 2006) or the overlapping group LASSO (Jacob et al., 2009), are also not suitable here, because such methods select *all* effects from an active group, whereas only a handful of effects may be active within a CME group.

In this light, a *bi-level* selection strategy is needed to select both *active CME groups* and *active effects within CME groups*. In recent years, there have been important developments on bi-level variable selection, including the sparse group LASSO (Wu and Lange, 2008; Simon et al., 2013) and the group exponential LASSO (Breheny and Huang, 2009; Breheny, 2015). We extend the latter framework here, because it allows us to encode within the penalization criterion two principles called *CME coupling* and *CME reduction*. These two principles guide the search for good CME models, and can be seen as an extension of effect

heredity and effect hierarchy (Wu and Hamada, 2009), two guiding principles used for model selection in designed experiments.

The paper is organized as follows. Section 2 provides some motivation for the problem at hand, including the implicit collinearity structure of CME groups and its effect on selection inconsistency. Section 3 proposes a new penalization criterion for CME selection, and introduces a coordinate descent optimization algorithm using threshold operators. Section 4 presents a tuning procedure for penalty parameters, and provides several tools for efficient optimization in high dimensions. Section 5 outlines several simulations comparing **cmenet** to existing methods. Section 6 then demonstrates the usefulness of the proposed technique in a gene association application, and Section 7 concludes with directions for future research.

2 Background and motivation

2.1 CME and CME groups

We first define some notation. Let $\mathbf{y} \in \mathbb{R}^n$ be a vector of n observations, and suppose p main effects are considered. For effect J , let $\tilde{\mathbf{x}}_j = (x_{1,j}, \dots, x_{n,j}) \in \{-1, +1\}^n$ be its binary covariate vector, $j = 1, \dots, p$. The tilde on $\tilde{\mathbf{x}}_j$ distinguishes the binary variable from its normalized analogue \mathbf{x}_j , which is introduced later. A CME can then be defined as follows:

Definition 1. *The conditional main effect (CME) of J given K at level $+$, denoted as $J|K+$, quantifies the effect of covariate vector $\tilde{\mathbf{x}}_{j|k+} = (\tilde{x}_{1,j|k+}, \dots, \tilde{x}_{n,j|k+})$, where:*

$$\tilde{x}_{i,j|k+} = \begin{cases} \tilde{x}_{i,j}, & \text{if } \tilde{x}_{i,k} = +1 \\ 0, & \text{if } \tilde{x}_{i,k} = -1 \end{cases}, \quad \text{for } i = 1, \dots, n.$$

The CME $J|K-$ can be defined in a similar manner.

A	B	$A B+$	$A B-$	$B A+$	$B A-$
+1	+1	+1	0	+1	0
+1	-1	0	+1	-1	0
-1	+1	-1	0	0	+1
-1	-1	0	-1	0	-1

Table 1: Model matrix for the two MEs A and B , and its four CMEs $A|B+$, $A|B-$, $B|A+$, $B|A-$.

Throughout this paper, the effects A and B are respectively referred to as the *parent effect* and the *conditioned effect* of $A|B+$. Using this terminology, $A|B+$ quantifies the effect of parent effect A , given its conditioned effect B is at level $+$. For illustration, Table 1 shows the four possible CMEs constructed from two main effects A and B .

Restricted to two-level, regular designed experiments, Su and Wu (2017) identified three important CME groups for selecting an *orthogonal* model, in which active predictors are orthogonal to each other. These three groups are: (a) *sibling* CMEs: CMEs with the same parent effect, (b) *twin* CMEs: CME pairs with the same parent and conditioned effect, but with the sign for the latter flipped, (c) *family* CMEs: CMEs with fully-aliased interaction effects. Leveraging the structure imposed by these groups, three simple rules were then proposed for selecting a parsimonious, orthogonal model.

Unfortunately, such groupings are not suitable for analyzing observational data, because an orthogonal model is most likely not attainable in this more general setting. Instead, by exploring the correlation structure of CMEs, the following new groups can be derived:

1. *Sibling* CMEs: CMEs with the same parent effect, e.g., $\{A|B+, A|C+, A|D+, \dots\}$. This is the same as in Su and Wu (2017),
2. *Cousin* CMEs: CMEs with the same conditioned effect, e.g., $\{B|A+, C|A+, D|A+, \dots\}$,
3. *Parent-child* pairs: An effect pair consisting of a CME and its parent ME, e.g.,

$$\{A|B+, A\}, \{A|C+, A\}, \dots,$$

4. *Uncle-nephew* pairs: An effect pair consisting of a CME and its conditioned ME, e.g., $\{B|A+, A\}, \{C|A+, A\}, \dots$.

We first outline the justification for these groups in terms of multi-collinearity, then discuss why such groupings are appealing from a selection consistency perspective.

2.2 Group structure for collinearity

To explore the group structure of CMEs, consider the following latent model for the main effects $\{\tilde{\mathbf{x}}_j\}_{j=1}^p \subseteq \{-1, +1\}^n$. Define the latent matrix $\mathbf{Z} = (z_{i,j})_{i=1}^n_{j=1}^p \in \mathbb{R}^{n \times p}$, where each row of \mathbf{Z} is drawn independently from the equicorrelated normal distribution $\mathcal{N}\{\mathbf{0}, \rho \mathbf{J}_p + (1 - \rho) \mathbf{I}_p\}$. Here, \mathbf{I}_p is the $p \times p$ identity matrix, \mathbf{J}_p is an $p \times p$ matrix of ones, and $\rho \in [0, 1]$. We then assume the following form for $\{\tilde{\mathbf{x}}_j\}_{j=1}^p$:

$$\tilde{x}_{i,j} = \mathbf{1}\{z_{i,j} > 0\} - \mathbf{1}\{z_{i,j} \leq 0\}, \quad i = 1, \dots, n, \quad j = 1, \dots, p. \quad (1)$$

Note that a larger value of ρ induces a higher correlation between the binary main effects.

Without loss of generality, assume here that the conditioned effects are set at level +1 for all CMEs. With the above model, the theorem below reveals an interesting group structure for CMEs. For brevity, proofs of technical results are deferred to the Appendix.

Theorem 1. *Under the latent model (1), the five effect pairs have the following correlations:*

<i>Effect pair</i>	<i>Correlation</i>	<i>Effect pair</i>	<i>Correlation</i>
Main effects	$\psi_{me}(\rho) = \frac{2 \sin^{-1} \rho}{\pi}$	Parent-child	$\psi_{pc}(\rho) = \frac{1}{2\sigma_c}$
Siblings	$\psi_{sib}(\rho) = \frac{1}{\sigma_c^2} \left\{ \frac{1}{4} + \frac{\sin^{-1} \rho}{2\pi} - \left(\frac{\sin^{-1} \rho}{\pi} \right)^2 \right\}$	Uncle-nephew	$\psi_{un}(\rho) = \frac{\sin^{-1} \rho}{\pi \sigma_c}$
Cousins	$\psi_{cou}(\rho) = \frac{1}{\sigma_c^2} \left\{ \frac{\sin^{-1} \rho}{\pi} - \left(\frac{\sin^{-1} \rho}{\pi} \right)^2 \right\}$		

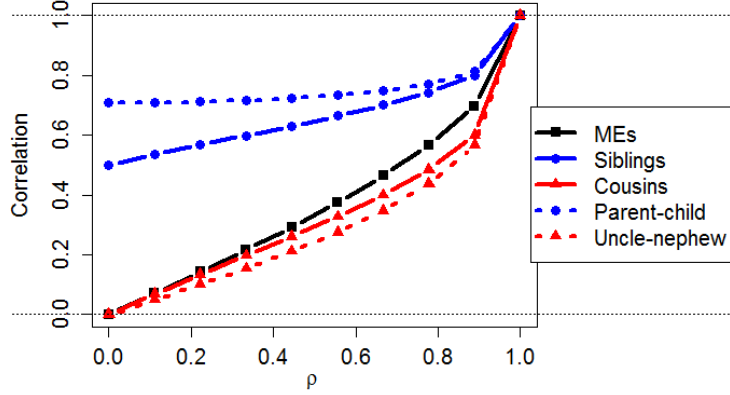


Figure 1: Average correlations of effect pairs as a function of ρ . This figure shows the highly varying correlations between MEs, siblings, cousins, parent-child pairs and uncle-nephew pairs.

where $\sigma_c^2 = 1/2 - (\sin^{-1} \rho/\pi)^2$.

Figure 1 plots the average correlations of the five effect pairs in Theorem 1. Two key observations can be made. First, the magnitude of these correlations impose a natural hierarchy for the effect groups. For all values of $\rho \in (0, 1)$, parent-child pairs and sibling pairs have the largest correlations, followed by ME pairs, then cousin and uncle-nephew pairs. Second, the relative correlation changes between groups can vary considerably for different choices of ρ . In the independent setting of $\rho = 0$, sibling and parent-child pairs exhibit high correlations of $1/2$ and $1/\sqrt{2}$, respectively, whereas the remaining three groups exhibit zero correlation. However, in the near-perfect correlation setting of $\rho \rightarrow 1^-$, the correlation for all groups converge to a common value of 1.

In light of this complex collinearity structure, one may suspect that standard variable selection techniques, such as the LASSO, would perform poorly for CME selection, because such methods impose the same regularization penalty over all predictors, thereby ignoring the grouped correlation structure implicit for CMEs. This is indeed the case, and we demonstrate this poor selection performance both in the following section and in the simulations of Section 5.

2.3 Selection inconsistency

An important property of a selection method is its *consistency* in choosing the correct model. Put mathematically, a method is (sign-) *selection consistent* if $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\beta}_n =_s \beta) = 1$, where $\beta \in \mathbb{R}^p$ is the true regression coefficient vector, $\hat{\beta}_n$ is the estimated vector using n observations, and $=_s$ denotes equality in sign (see Zhao and Yu, 2006 for a precise definition). The following theorem shows that LASSO is inconsistent for very simple CME models:

Theorem 2. *Using the model in (1), the LASSO is selection inconsistent in the following situations: (a) for $\rho \geq 0$, a model with $q \geq 3$ active siblings, (b) for $\rho \geq 0.27$, a model with $q = 2$ active main effects, and (c) for $\rho \geq 0.29$, a model with $q \geq 6$ active cousins.*

The proof of this relies on the *irrepresentability condition* in Zhao and Yu (2006), which shows that the LASSO is selection inconsistent when active variables are highly correlated with non-active ones. In view of the highly correlated grouped structure in CME models, Theorem 2 shows that the LASSO provides poor selection performance for basic CME models, even when little to no correlation is present for the underlying main effects.

3 cmenet: a new method for CME selection

To address the above problems, we propose a novel bi-level variable selection method called **cmenet**, which can identify both active CME groups and active effects within such groups. Similar to popular selection methods such as the elastic net (Zou and Hastie, 2005) and SparseNet (Mazumder et al., 2012), the name **cmenet** draws an analogy between the proposed method’s ability to select the active variables amongst non-active ones and a fishing net’s ability to catch the larger fish amongst smaller ones. **cmenet** features two important principles, called *CME coupling* and *CME reduction*, which guide the selection procedure. This section concludes with an optimization algorithm which exploits coordinate

descent, majorization-minimization and threshold operators for efficient model selection.

3.1 Selection criterion

We now introduce the selection criterion. Let $\mathbf{x}_j \in \mathbb{R}^n$ be the normalization of the main effect $\tilde{\mathbf{x}}_j$, with $\mathbf{x}_j^T \mathbf{1}_n = 0$ and $\|\mathbf{x}_j\|_2 = 1$ (i.e., zero mean and unit norm), with a similar notation for CMEs. Further let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{p'}) \in \mathbb{R}^{n \times p'}$ be the full model matrix consisting of the normalized ME and CME effects, where $p' = p + 4\binom{p}{2}$ is the total number of effects considered. For simplicity, assume all predictors are binary for the following exposition; Section 3.4 gives a simple extension to the case of both binary and continuous predictors. Let $\boldsymbol{\beta} \in \mathbb{R}^{p'}$ be the regression coefficient vector, with β_j and $\beta_{j|k+}$ its corresponding coefficients for ME J and CME $J|K+$. Finally, assume that \mathbf{y} is centered, i.e., $\mathbf{y}^T \mathbf{1}_n = 0$.

For effect groups, define $\mathcal{S}(j) = \{J, J|A+, J|A-, J|B+, J|B-, \dots\}$ as the *sibling group* for parent effect j , and $\mathcal{C}(j) = \{J, A|J+, A|J-, B|J+, B|J-, \dots\}$ as the *cousin group* for conditioned effect j , $j = 1, \dots, p$. We propose the following selection criterion:

$$\begin{aligned} \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) &\equiv \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_{\mathcal{S}}(\boldsymbol{\beta}) + P_{\mathcal{C}}(\boldsymbol{\beta}) \right\}, \\ P_{\mathcal{S}}(\boldsymbol{\beta}) &\equiv \sum_{j=1}^p f_{o,\mathcal{S}} \left\{ \sum_{k \in \mathcal{S}(j)} f_{i,\mathcal{S}}(\beta_k) \right\}, \quad P_{\mathcal{C}}(\boldsymbol{\beta}) \equiv \sum_{j=1}^p f_{o,\mathcal{C}} \left\{ \sum_{k \in \mathcal{C}(j)} f_{i,\mathcal{C}}(\beta_k) \right\}. \end{aligned} \quad (2)$$

Here, $f_{o,\mathcal{S}}$ and $f_{i,\mathcal{S}}$ (similarly, $f_{o,\mathcal{C}}$ and $f_{i,\mathcal{C}}$) are *outer* and *inner* penalties which control the *between-group* and *within-group* selection for sibling (similarly, cousin) groups, respectively. This can be seen as an extension of the hierarchical framework in Breheny and Huang (2009). While the specific penalty functions are left arbitrary in (2), we will introduce **cmenet** for the specific choice of the exponential penalty in Breheny (2015) for outer penalty, and the

minimax concave-plus penalty (MC+) in Zhang (2010) for inner penalty:

$$\begin{aligned} \text{Outer: } f_{o,S}(\theta) &= \eta_{\lambda_s, \tau}(\theta), \quad f_{o,C}(\theta) = \eta_{\lambda_c, \tau}(\theta), \quad \text{where } \eta_{\lambda, \tau}(\theta) = \frac{\lambda^2}{\tau} \left\{ 1 - \exp\left(-\frac{\tau\theta}{\lambda}\right) \right\}, \\ \text{Inner: } f_{i,S}(\beta) &= g_{\lambda_s, \gamma}(\beta), \quad f_{i,C}(\beta) = g_{\lambda_c, \gamma}(\beta), \quad \text{where } g_{\lambda, \gamma}(\beta) = \int_0^{|\beta|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx. \end{aligned} \quad (3)$$

The appeal for the so-called exponential-MC+ framework above is that it provides a concise parametrization of the grouped collinearity structure in Section 2. First, the penalties $\lambda_s > 0$ and $\lambda_c > 0$ allow for differing regularization levels within sibling and cousin groups, respectively, with larger penalties reducing the number of selected effects in each group. Assuming such parameters are tuned using cross-validation, a smaller tuned value of λ_s suggests many sibling effects are present in the data, while a smaller λ_c suggests the same for cousin effects. Second, the parameter $\gamma > 1$ controls the non-convexity of the inner MC+ penalty, and provides a “bridge” between the l_0 -penalty (obtained when $\gamma \rightarrow 1^+$) and the l_1 -penalty in LASSO (obtained when $\gamma \rightarrow \infty$). In view of the selection problems for LASSO (see Theorem 2), such a parameter allows for improved selection performance of the highly correlated CMEs. Lastly, the parameter τ provides two appealing features called CME coupling and reduction, which we introduce below.

3.2 CME coupling and reduction

Consider first a CME $J|K+$ which has yet to be selected (i.e., $\beta_{j|k+} = 0$), and assume without loss of generality that $\mathbf{x}_{j|k+}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n > 0$. Taking the derivative of $Q(\boldsymbol{\beta})$ with respect to $\beta_{j|k+}$, we get:

$$\begin{aligned} \frac{\partial}{\partial \beta_{j|k+}} Q(\boldsymbol{\beta}) \Big|_{\beta_{j|k+}=0} &= -\frac{1}{n} \mathbf{x}_{j|k+}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \Delta_{S(j)} + \Delta_{C(k)}, \\ \text{where } \Delta_{S(j)} &= \lambda_s \exp \left\{ -\frac{\tau \|\boldsymbol{\beta}_{S(j)}\|_{\lambda_s, \gamma}}{\lambda_s} \right\} \quad \text{and} \quad \Delta_{C(k)} = \lambda_c \exp \left\{ -\frac{\tau \|\boldsymbol{\beta}_{C(k)}\|_{\lambda_c, \gamma}}{\lambda_c} \right\}. \end{aligned} \quad (4)$$

Here, $\beta_g \in \mathbb{R}^{|g|}$ denotes the coefficient vector for an effect subset $g \subseteq \{1, \dots, p'\}$, and $\|\beta_g\|_{\lambda, \gamma} \equiv \sum_{l \in g} g_{\lambda, \gamma}(\beta_l)$ denotes its “norm” under the inner MC+ penalty. Note that, when more and more effects have been selected in the sibling group $\mathcal{S}(j)$ (or cousin group $\mathcal{C}(k)$), the linearized slope $\Delta_{\mathcal{S}(j)}$ (or $\Delta_{\mathcal{C}(k)}$) becomes smaller and smaller, which in turn causes a decrease in the derivative $\frac{\partial}{\partial \beta_{j|k+}} Q(\beta)$. Since the goal is to minimize the selection criterion $Q(\beta)$, a smaller derivative thereby allows for greater decrease in $Q(\beta)$ when $\beta_{j|k+}$ enters the model. In other words, the CME $J|K+$ has a greater chance of entering the model when other effects in its sibling group $\mathcal{S}(j)$ or its cousin group $\mathcal{C}(k)$ have already been selected. We call this feature *CME coupling*, following the idea of effect coupling in Breheny (2015).

Consider next a ME J which has yet to be selected (i.e., $\beta_j = 0$), and assume again that $\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\beta)/n > 0$. The derivative of $Q(\beta)$ for β_j becomes:

$$\left. \frac{\partial}{\partial \beta_j} Q(\beta) \right|_{\beta_j=0} = -\frac{1}{n} \mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\beta) + \Delta_{\mathcal{S}(j)} + \Delta_{\mathcal{C}(j)}. \quad (5)$$

When more and more effects have already been selected in the sibling group $\mathcal{S}(j)$ (or the cousin group $\mathcal{C}(j)$), the linearized slopes $\Delta_{\mathcal{S}(j)}$ (or $\Delta_{\mathcal{C}(j)}$) become smaller and smaller, thereby decreasing the derivative $\frac{\partial}{\partial \beta_j} Q(\beta)$ in (5). Hence, the main effect J enters the model more easily when effects in its sibling group $\mathcal{S}(j)$ or its cousin group $\mathcal{C}(j)$ have already been selected. We refer to this phenomenon as *CME reduction*.

The notions of CME coupling and reduction are quite intuitive to expect for many CME applications. Consider the gene expression example in the Introduction, where selection of the CME $A|B+$ indicates the effectiveness of gene A only when gene B is present. When several siblings of A , say, $A|B+$ and $A|C+$, are already selected in the model, one naturally expects sibling effects for A to be more likely active than effects with no selected siblings, which is precisely CME coupling. However, when many sibling effects of A have already been selected, one may suspect that the main effect A is active instead of the selected

siblings, which is precisely CME reduction. A similar analogy holds for cousin effects.

An interesting parallel can also be made connecting CME coupling and reduction with the two guiding principles for model selection in designed experiments (Wu and Hamada, 2009). The first principle, called (weak) *effect heredity*, states that higher-order interactions can be selected only when either of its parent main effects are in the model. This idea is quite similar to coupling, which allows for easier selection of a CME when effects with either the same parent or conditioned ME have been selected. Furthermore, note that a CME can be interpreted as a component of an interaction effect, because the difference of the two CMEs $A|B+$ and $A|B-$ is the interaction $A * B$ (Su and Wu, 2017). Coupling can therefore be seen as an *extension* of effect heredity after breaking an interaction effect (which is often difficult to interpret) into more interpretable conditional effects. The second principle, called *effect hierarchy*, states that lower-order interactions are more likely active than higher-order ones. This is akin to CME reduction, which encourages the reduction of selected sibling (or cousin) CMEs to its parent (conditioned) effect when too many siblings (cousins) are in the model.

3.3 Coordinate descent and threshold operators

We now develop the technical framework for minimizing the selection criterion $Q(\boldsymbol{\beta})$. A key tool in the optimization algorithm is coordinate descent, which can be explained as follows. Viewing $Q(\boldsymbol{\beta})$ as a function of only the first coefficient β_1 (call this $Q_1(\beta_1)$), we first update β_1 as the minimizer of $Q_1(\cdot)$, keeping the remaining $p' - 1$ coefficients fixed. The same procedure is then applied cyclically over $\beta_2, \dots, \beta_{p'}$, and repeated until the full coefficient vector $\boldsymbol{\beta}$ converges. In recent years, coordinate descent has become widely used in the variable selection literature (see, e.g., Fu, 1998; Friedman et al., 2007; Mazumder et al., 2011), due to its simplicity and efficiency for high dimensional problems. The key to efficiency lies in the existence of a *closed-form* minimizer for the coordinate-wise objective

$Q_j(\cdot)$, also known as a *threshold function* from signal processing (Donoho, 1995). We derive below such a threshold function for $Q(\boldsymbol{\beta})$.

Before delving into details, we first investigate the convexity properties of $Q(\boldsymbol{\beta})$:

Proposition 1. *$Q(\boldsymbol{\beta})$ is strictly convex when $\tau < 1/(2n)\lambda_{\min}(\mathbf{X}^T\mathbf{X})$, where each column of \mathbf{X} is normalized with zero mean and unit norm, and $\lambda_{\min}(\cdot)$ returns the minimum eigenvalue. Also, for any $j = 1, \dots, p'$, $Q_j(\beta_j)$ is strictly convex when $\tau < 1/(2n)$.*

In words, this shows that a sufficiently small choice of τ is needed to ensure some form of convexity for the objective $Q(\boldsymbol{\beta})$. The first part of this proposition shows a unique global minimum exists for $Q(\boldsymbol{\beta})$ when $\tau < 1/(2n)\lambda_{\min}(\mathbf{X}^T\mathbf{X})$. Such a result is quite restrictive, because it applies only to the low-dimensional setting of $n \leq p'$, where $\lambda_{\min}(\mathbf{X}^T\mathbf{X})$ is strictly positive. The second part guarantees the coordinate-wise objective $Q_j(\beta_j)$ is strictly convex when $\tau < 1/(2n)$, a result which holds in the high-dimensional setting of $n > p'$. This coordinate-wise convexity is important for deriving the threshold function below.

For a main effect J , consider its coordinate-wise minimization:

$$\min_{\beta_j} Q_j(\beta_j) = \min_{\beta_j} \left[\frac{1}{2n} \|\mathbf{r}_{-j} - \mathbf{x}_j \beta_j\|_2^2 + f_{\lambda_s, \tau} \{ \|\boldsymbol{\beta}_{\mathcal{S}(j)}\|_{\lambda_s, \gamma} \} + f_{\lambda_c, \tau} \{ \|\boldsymbol{\beta}_{\mathcal{C}(j)}\|_{\lambda_c, \gamma} \} \right], \quad (6)$$

where $\mathbf{r}_j = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_j \beta_j$ is the residual vector fitted without \mathbf{x}_j . Similarly, for a CME $J|K+$, its coordinate-wise minimization becomes:

$$\min_{\beta_{j|k+}} Q_{j|k+}(\beta_{j|k+}) = \min_{\beta_{j|k+}} \left[\frac{1}{2n} \|\mathbf{r}_{-(j|k+)} - \mathbf{x}_{j|k+} \beta_{j|k+}\|_2^2 + f_{\lambda_s, \tau} \{ \|\boldsymbol{\beta}_{\mathcal{S}(j)}\|_{\lambda_s, \gamma} \} + f_{\lambda_c, \tau} \{ \|\boldsymbol{\beta}_{\mathcal{C}(k)}\|_{\lambda_c, \gamma} \} \right]. \quad (7)$$

An optimization technique called *majorization-minimization* (MM, see Chapter 12 of Lange, 2010) can now be used to derive a threshold function. The main idea of MM is as follows. Instead of minimizing the original objective function, one first obtains a majorizing surrogate which lies above the objective, then minimizes this surrogate function instead. Under

certain conditions, the solution iterates generated by repeating this procedure converge to a minimizer for the original problem (Lange, 2010). For Q_j and $Q_{j|k+}$, a simple first-order expansion provides a nice majorizing surrogate which can be minimized in closed form, as the following theorem demonstrates.

Theorem 3. Suppose $\tau < 1/(2n)$. For fixed $\tilde{\beta} \in \mathbb{R}^{p'}$, define $\bar{Q}_j(\cdot|\tilde{\beta})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\beta})$ as:

$$\begin{aligned}\bar{Q}_j(\beta_j|\tilde{\beta}) &= \frac{1}{2n} \|\tilde{\mathbf{r}}_{-j} - \mathbf{x}_j \beta_j\|_2^2 + P_s(\tilde{\beta}) + P_c(\tilde{\beta}) \\ &\quad + \tilde{\Delta}_{\mathcal{S}(j)} \left\{ g_{\lambda_s, \gamma}(\beta_j) - g_{\lambda_s, \gamma}(\tilde{\beta}_j) \right\} + \tilde{\Delta}_{\mathcal{C}(j)} \left\{ g_{\lambda_c, \gamma}(\beta_j) - g_{\lambda_c, \gamma}(\tilde{\beta}_j) \right\}, \text{ and} \\ \bar{Q}_{j|k+}(\beta_{j|k+}|\tilde{\beta}) &= \frac{1}{2n} \|\tilde{\mathbf{r}}_{-(j|k+)} - \mathbf{x}_{j|k+} \beta_{j|k+}\|_2^2 + P_s(\tilde{\beta}) + P_c(\tilde{\beta}) \\ &\quad + \tilde{\Delta}_{\mathcal{S}(j)} \left\{ g_{\lambda_s, \gamma}(\beta_{j|k+}) - g_{\lambda_s, \gamma}(\tilde{\beta}_{j|k+}) \right\} + \tilde{\Delta}_{\mathcal{C}(k)} \left\{ g_{\lambda_c, \gamma}(\beta_{j|k+}) - g_{\lambda_c, \gamma}(\tilde{\beta}_{j|k+}) \right\},\end{aligned}$$

where $\tilde{\cdot}$ indicates the quantity is computed with $\tilde{\beta}$ instead of β . Then:

- a) $\bar{Q}_j(\cdot|\tilde{\beta})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\beta})$ are majorization functions for $Q_j(\cdot)$ and $Q_{j|k+}(\cdot)$, respectively,
- b) The unique minimizers of $\bar{Q}_j(\cdot|\tilde{\beta})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\beta})$ are given by $S_{\lambda_s, \lambda_c}(\mathbf{x}_j^T \mathbf{r}_{-j}/n; \tilde{\Delta}_{\mathcal{S}(j)}, \tilde{\Delta}_{\mathcal{C}(j)})$ and $S_{\lambda_s, \lambda_c}(\mathbf{x}_{j|k+}^T \mathbf{r}_{-j|k+}/n; \tilde{\Delta}_{\mathcal{S}(j)}, \tilde{\Delta}_{\mathcal{C}(k)}),$ respectively.

Here, $S_{\lambda_1, \lambda_2}(\cdot; \Delta_1, \Delta_2)$ is the threshold function:

$$S_{\lambda_1, \lambda_2}(z; \Delta_1, \Delta_2) = \begin{cases} z & \text{if } z \in [\lambda_{(1)}\gamma, \infty), \\ \text{sgn}(z) (|z| - \Delta_{(1)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma}\right) & \\ & \text{if } z \in \left[\lambda_{(2)}\gamma + \Delta_{(1)} \left(1 - \frac{\lambda_{(2)}}{\lambda_{(1)}}\right), \lambda_{(1)}\gamma\right), \\ \text{sgn}(z) (|z| - \Delta_{(1)} - \Delta_{(2)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma} - \frac{\Delta_{(2)}}{\lambda_{(2)}\gamma}\right) & \\ & \text{if } z \in \left[\Delta_{(1)} + \Delta_{(2)}, \lambda_{(2)}\gamma + \Delta_{(1)} \left(1 - \frac{\lambda_{(2)}}{\lambda_{(1)}}\right)\right), \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

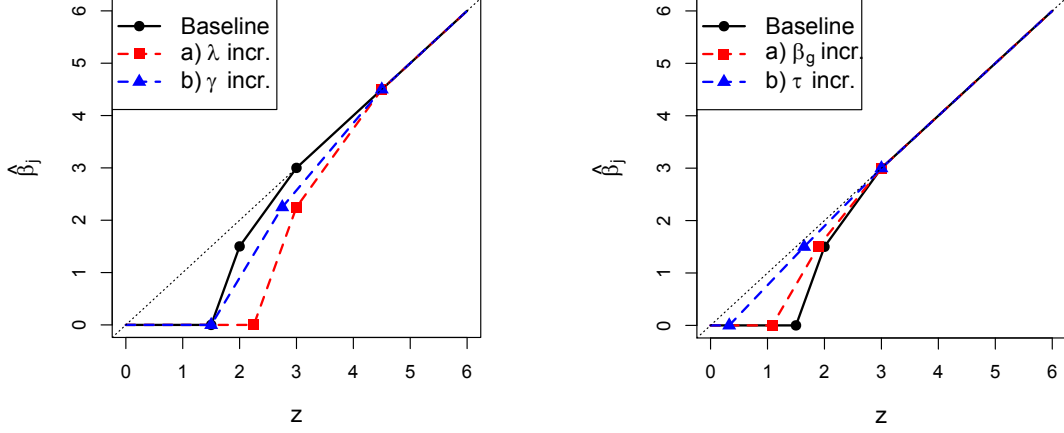


Figure 2: Plots of the threshold function S_{λ_1, λ_2} for various parameter settings. The baseline setting is fixed at $(\lambda_1, \lambda_2, \gamma, \tau) = (1, 0.5, 3, 0.05)$ with no selected group effects. (Left) Threshold functions for the baseline setting and two new settings $(1.5, 0.75, 3, 0.05)$ and $(1, 0.5, 4.5, 0.05)$, all with no selected group effects. (Right) Threshold functions for the baseline setting, and two new settings $(1, 0.5, 3, 0.05)$ and $(1, 0.5, 3, 0.25)$ with grouped norms $\|\beta_g\|_{\lambda_1, \gamma} = \|\beta_g\|_{\lambda_2, \gamma} = 5$.

where $\lambda_{(1)} = \max(\lambda_1, \lambda_2)$ and $\lambda_{(2)} = \min(\lambda_1, \lambda_2)$, with $\Delta_{(1)}$ and $\Delta_{(2)}$ its corresponding slopes.

To better understand the shrinkage behavior of this threshold function, Figure 2 plots $S_{\lambda_1, \lambda_2}(z; \Delta_1, \Delta_2)$ for various choices of $\lambda_1, \lambda_2, \gamma, \tau$ and grouped norms $\|\beta_g\|_{\lambda_1, \gamma}$ and $\|\beta_g\|_{\lambda_2, \gamma}$. Consider first the baseline setting of $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\gamma = 3$ and $\tau = 0.05$, with $\|\beta_g\|_{\lambda_1, \gamma} = \|\beta_g\|_{\lambda_2, \gamma} = 0$ (i.e., no selected grouped effects). From the plot, the threshold function can be seen to be continuous and piecewise linear in four segments. The first segment is a horizontal line at zero, and represents the values for which a coefficient is shrunk to zero after thresholding. The last segment, which is the identity line, corresponds to values for which the full coefficient signal is retained without any shrinkage. The middle two segments provide a two-step transition between these two extremes, with slopes controlled by both sibling and cousin penalties. This is similar to the threshold function for the MC+ penalty, except the latter achieves this transition in one step. The two-step transition in

Algorithm 1 cmenet: An algorithm for bi-level CME selection

```

1: function CMENET( $\mathbf{X}, \mathbf{y}, \lambda_s, \lambda_c, \gamma, \tau, \boldsymbol{\beta} = \mathbf{0}_{p'}$ )
2:   • Initialize  $\mathbf{r} \leftarrow \mathbf{y} - \bar{\mathbf{y}}$ ,  $\Delta_{\mathcal{S}(j)} = \lambda_s$ ,  $\Delta_{\mathcal{C}(j)} = \lambda_c$  for  $j = 1, \dots, p$ 
3:   repeat
4:     for  $j = 1, \dots, p$  do ▷ For all main effects...
5:       •  $\beta_0 \leftarrow \beta_j$ ,  $\beta_j \leftarrow S_{\lambda_s, \lambda_c}\{(\mathbf{x}_j^T \mathbf{r} + \beta_0)/n; \Delta_{\mathcal{S}(j)}, \Delta_{\mathcal{C}(j)}\}$  ▷ Shrinkage
6:       •  $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{x}_j(\beta_0 - \beta_j)$  ▷ Update residual
7:       •  $\Delta_{\mathcal{S}(j)} \leftarrow \Delta_{\mathcal{S}(j)} \exp\{-\tau/\lambda_s [g_{\lambda_s, \gamma}(\beta_j) - g_{\lambda_s, \gamma}(\beta_0)]\}$  ▷ Update slopes
8:       •  $\Delta_{\mathcal{C}(j)} \leftarrow \Delta_{\mathcal{C}(j)} \exp\{-\tau/\lambda_c [g_{\lambda_c, \gamma}(\beta_j) - g_{\lambda_c, \gamma}(\beta_0)]\}$ 
9:     for  $j = 1, \dots, p$  and  $k = 1, \dots, p$  do ▷ For all CMEs (both  $J|K+$  and  $J|K-$ ) ...
10:      •  $\beta_0 \leftarrow \beta_{j|k+}$ ,  $\beta_{j|k+} \leftarrow S_{\lambda_s, \lambda_c}\{(\mathbf{x}_{j|k+}^T \mathbf{r} + \beta_0)/n; \Delta_{\mathcal{S}(j)}, \Delta_{\mathcal{C}(k)}\}$  ▷ Shrinkage
11:      •  $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{x}_{j|k+}(\beta_0 - \beta_{j|k+})$  ▷ Update residual
12:      •  $\Delta_{\mathcal{S}(j)} \leftarrow \Delta_{\mathcal{S}(j)} \exp\{-\tau/\lambda_s [g_{\lambda_s, \gamma}(\beta_{j|k+}) - g_{\lambda_s, \gamma}(\beta_0)]\}$  ▷ Update slopes
13:      •  $\Delta_{\mathcal{C}(k)} \leftarrow \Delta_{\mathcal{C}(k)} \exp\{-\tau/\lambda_c [g_{\lambda_c, \gamma}(\beta_{j|k+}) - g_{\lambda_c, \gamma}(\beta_0)]\}$ 
14:   until  $\boldsymbol{\beta}$  converges
   return the converged coefficient vector  $\boldsymbol{\beta}$ 

```

Figure 2 is quite appealing for CME selection, because it implicitly controls the two-tiered coupling effect from sibling and cousin groups.

Next, consider the varying parameter settings in the left and right plots of Figure 2. From the left plot, an increase in λ_1 , λ_2 or γ appears to cause greater shrinkage of the coefficient signal. This is expected, because a larger λ_1 and λ_2 induces greater regularization, and a larger γ generates a “more convex” penalty function. From the right plot, an increase in τ in the presence of selected group effects appears to greatly reduce the shrinkage applied to a coefficient. This is precisely the notion of CME coupling, where the selection of sibling or cousin effects greatly increases the chances of a CME entering the model.

3.4 Algorithm statement

Putting all the pieces together, Algorithm 1 summarizes the full steps for minimizing the selection criterion (2) given fixed parameters λ_s , λ_c , γ and τ . Starting with an initial

solution of $\beta = \mathbf{0}_{p'}$, the threshold function in (8) is applied cyclically for each entry of β . This iterative procedure is then repeated until β converges. Using the majorization function in Theorem 3, one can prove the convergence of **cmenet** to a stationary solution.

Corollary 1. *When $\tau < 1/(2n)$, **cmenet** converges to a solution $\hat{\beta}$ satisfying $\nabla Q(\hat{\beta}) = 0$.*

While the above exposition for **cmenet** considers only the selection of CMEs, the proposed method can easily be *extended to the selection of both CMEs and continuous predictors*. Assuming the l_1 -penalty in LASSO is used for the latter, one can simply apply the soft-thresholding operator (Donoho, 1995) to the coefficients of each continuous variable within the coordinate descent loop in Algorithm 1. The algorithmic convergence for such a modification is analogous to Corollary 1, and is not included for brevity.

Lastly, regarding the running time of **cmenet**, one can show that one coordinate descent cycle over all p' coefficients requires $\mathcal{O}(np')$ time, because each coordinate descent step requires $\mathcal{O}(n)$ work. The linear running time in both n and p' is crucial for the computational efficiency of **cmenet** in the high dimensional setting of $p \gg 1$.

4 cv.cmenet and computational speed ups

In this section, we first propose an algorithm called **cv.cmenet** for tuning the selection parameters λ_s , λ_c , γ and τ in **cmenet**. Several computational tools are then introduced which allows **cv.cmenet** to be highly efficient in high-dimensions. These include warm start initialization, active set optimization and strong rules for screening inactive effects.

4.1 cv.cmenet: an algorithm for tuning cmenet

Since four selection parameters are considered here, we first provide some guiding rules for efficiently exploring the parameter space. The following proposition provides one such rule

for λ_s and λ_c :

Proposition 2. *Suppose $\lambda_s + \lambda_c \geq \max_{j=1, \dots, p'} |\mathbf{x}_j^T \mathbf{y}|$. When $Q(\boldsymbol{\beta})$ is strictly convex, the unique minimizer of $Q(\boldsymbol{\beta})$ is the trivial solution $\boldsymbol{\beta} = \mathbf{0}_{p'}$.*

In the high dimensional setting of $n > p'$, $Q(\boldsymbol{\beta})$ cannot be strictly convex (see discussion for Proposition 1), so $\boldsymbol{\beta} = \mathbf{0}_{p'}$ is only a stationary solution. Nonetheless, such a rule allows for considerable reduction in the search for interesting choices of λ_s and λ_c . From Proposition 1, another rule is $\tau < 1/(2n)$, which ensures the coordinate-wise problem is strictly convex.

After choosing a feasible set of parameters, a technique called K -fold cross validation (CV) (Friedman et al., 2001) is then used to estimate the prediction error for that parameter combination. The idea of K -fold CV can be explained as follows. First, the data is randomly split into K equal-sized parts, with $K - 1$ parts used for model training using `cmenet`, and the remaining part used to validate the trained model. An estimate of predictive error can then be obtained by repeating this validation procedure multiple times and averaging the resulting realizations of the loss function. The goal is to estimate the predictive errors at a large number of parameter settings, then pick the setting with the smallest error estimate.

Summarizing the above, the steps for `cv.cmenet` are as follows. First, the sequences $(\lambda_l)_{l=1}^L$, $\lambda_l < \max_{j=1, \dots, p'} |\mathbf{x}_j^T \mathbf{y}|$, $(\gamma_l)_{l=1}^G$, $\gamma_l > 0$ and $(\tau_l)_{l=1}^T$, $\tau_l < 1/(2n)$ are used to generate a grid of feasible parameter choices. Next, initial choices of (λ_s, λ_c) are set using the parameter setting with lowest predictive error for the MC+ penalty (in our implementation, this is estimated using K -fold CV through the R package `SPARSENET`). The parameters (γ, τ) and (λ_s, λ_c) are then iteratively optimized by picking the setting with lowest error from K -fold CV. Finally, the selected model is obtained using the optimal parameter combination. Details on this procedure are outlined in Algorithm 2 of the Appendix.

4.2 Warm starts, active sets and variable screening

With the help of three computational techniques, the proposed method `cv.cmenet` can be made much more efficient in high dimensions. The first technique, called warm starts, uses the converged solution from a previous parameter setting as the initial solution for the current setting. The use of warm starts in variable selection was popularized in Friedman et al. (2007), where the authors demonstrated its usefulness in efficiently optimizing coefficients along the full LASSO path. Implicit in this method is the assumption that the coefficient path is sufficiently smooth over the space of tuning parameters. For `cv.cmenet`, warm starts are implemented in lines 9 - 10 and 17 - 18 of Algorithm 2 in the Appendix.

The second technique, called active set optimization, is also a popular method used in many variable selection algorithms (e.g., Meier et al., 2008; Friedman et al., 2010). For `cmenet`, active set optimization simply means the coordinate descent updates are cycled over a small subset of *active* variables, instead of over the full set of p' variables. Such a technique is important for computational efficiency, since we know that the minimizer of $Q(\beta)$ is sparse by construction. In our implementation, the active set is determined by initially performing the full coordinate descent cycle for 25 iterations. Following Friedman et al. (2010), after the active set iterations converge, a full cycle is then performed over all p' variables. If this cycle does not change the active set, `cmenet` is terminated; otherwise the active set is updated, and the process repeated.

When p grows large, however, the full cycles needed to set the initial active set can be very time-consuming. The last technique, called variable screening, addresses this by initially screening a large number of effects using the so-called *strong rules*. Such rules are first termed in Tibshirani et al. (2012), where the authors used previous solved coefficient solutions along the LASSO path to screen out inactive variables for the current penalty. We employ a similar technique here. Fixing λ_c , γ and τ , let $\hat{\beta}(\lambda_s) \in \mathbb{R}^p$ be an optimal solution

of $Q(\beta)$ under sibling penalty λ_s . Furthermore, let $0 < \lambda_s^1 < \dots < \lambda_s^L < \max_{j=1, \dots, p'} |\mathbf{x}_j^T \mathbf{y}|$ be the penalty sequence for λ_s . The following theorem provides a strong rule for screening inactive effects in $\hat{\beta}(\lambda_s^l)$ using the previous solution $\hat{\beta}(\lambda_s^{l+1})$:

Theorem 4. *For fixed λ_c , γ and τ , define $c_j(\lambda_s) = \mathbf{x}_j^T (\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_s))/n$ for an effect j , and suppose effect j , along with its sibling and cousin groups \mathcal{S} and \mathcal{C} , are inactive at penalty λ_s^{l+1} . Assume the following:*

A.1 $nc_j(\lambda_s)$ is non-expansive in λ_s : $|c_j(\lambda_s) - c_j(\lambda'_s)| \leq |\lambda_s - \lambda'_s|/n \quad \forall \lambda_s \neq \lambda'_s$,

A.2 $\overline{\Delta\beta}_l \equiv \overline{\nabla\beta}(\lambda_s^{l+1} - \lambda_s^l) \leq \gamma \min(\lambda_s^l, \lambda_c)$, where $\overline{\nabla\beta} \equiv (2 + \frac{1}{n} + \frac{\tau\gamma}{2})(\frac{2}{\gamma} \exp\{-\frac{\tau\gamma}{2}\})$,

A.3 No effects from \mathcal{S} and \mathcal{C} are active under penalty λ_s^l .

Define $M_l = -(\lambda_s^{l+1} - \lambda_s^l)/n + D(\overline{\Delta\beta}_l; \lambda_s^l) (1 - \overline{\Delta\beta}_l/(\lambda_s^l \gamma)) + D(\overline{\Delta\beta}_l; \lambda_c) (1 - \overline{\Delta\beta}_l/(\lambda_c \gamma))$, and $D(\beta; \lambda) = \lambda \exp\{-(\tau/\lambda)g_{\lambda, \gamma}(\beta)\}$. If $|c_j(\lambda_s^{l+1})| < M_l$, then $\hat{\beta}_j(\lambda_s^l)$ must equal 0.

While Theorem 4 may seem complicated, the underlying idea is straight-forward. For an inactive effect j at previous penalty λ_s^{l+1} , the goal is to use $c_j(\lambda_s^{l+1})$, its inner product with the previous residual, to determine whether j will be inactive at the current penalty λ_s^l . When this inner product is smaller than cut-off M_l , effect j can be safely discarded from the active set; when it does exceeds M_l , j should be marked as active. In simulation studies with $p' = 50 + 4\binom{50}{2}$ total effects, this screening rule is able to correctly discard 97% of inactive effects on average, which greatly speeds up `cv.cmenet`. Such a rule is only needed over (λ_s, λ_c) space, because from simulations, the tuning over (γ, τ) space appears to be sufficiently fast using only warm starts and active sets.

We briefly provide some context for the assumptions in Theorem 4. A.1 is a typical smoothness assumption for the inner product path in LASSO (Tibshirani et al., 2012), A.2 is a similar smoothness assumption on the coefficient path $\hat{\beta}_j(\lambda_s)$, and A.3 assumes effect j is the first to enter the model in its sibling and cousin groups. It is worth emphasizing that

such assumptions *need not be always satisfied* in practice, because the screening rule is used as a *heuristic* for initializing the active set. As before, a full cycle over all p' variables can be performed to ensure the converged solution is indeed stationary.

5 Simulations

Here, we explore the performance of the proposed method in several simulation studies. The set-up is as follows. A total of 48 simulation cases are considered here, with varying choices of main effects p , number of active groups, number of active effects within a group, and whether the grouped effects are siblings, cousins or main effects. Active effects are assigned a value of 1 in β , and non-active effects assigned a value of 0. Each simulation case is then replicated 500 times, with the model matrix \mathbf{X} simulated from the equicorrelated latent model in Section 2.2 with $\rho = 0$ and $\rho = 1/\sqrt{2}$, and the response \mathbf{y} simulated independently from $\mathcal{N}(\mathbf{X}\beta, \mathbf{I}_n)$. For brevity, only results for $n = 20$ are reported here, but similar results hold for varying choices of n . Table 2 summarizes the above settings.

Under this set-up, we compare the proposed method `cv.cmenet` with two popular variable selection techniques from the literature: the LASSO (Tibshirani, 1996) using the R package GLMNET (Friedman et al., 2009), and SparseNet (Mazumder et al., 2011) using the R package SPARSENET (Mazumder et al., 2012). All three methods perform selection on the same set of ME and CME effects, with parameters tuned using 10-fold CV. Two criteria are used for comparison. The first criterion is the number of misspecified variables: $\#\{\mathcal{A} \setminus \hat{\mathcal{A}}_n\} + \#\{\hat{\mathcal{A}}_n \setminus \mathcal{A}\}$, where \mathcal{A} is the true active set and $\hat{\mathcal{A}}_n$ is the set of selected effects after n observations. Smaller values of this indicate better selection performance of a method. The second criterion is the mean-squared prediction error (MSPE): $\mathbb{E}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2$, with smaller values suggesting better predictive performance for a method.

For latent correlations of $\rho = 0$ and $\rho = 1/\sqrt{2}$, Figures 3 and 4 show the number of

<i>Simulation parameters</i>	<i>Settings</i>
# of main effects considered (total effects considered)	$p = 25$ or 50 $(p' = 25 + 4\binom{25}{2} \text{ or } 50 + 4\binom{50}{2})$
# of active groups ¹	2 or 4
# of active effects within a group ²	4 or 8
Effect type	Siblings, cousins, main effects
Latent correlation in model matrix	$\rho = 0$ or $1/\sqrt{2}$

Table 2: *Parameter settings for simulation study.*

¹ For effect type main effect, this is the # of active MEs.

² Not used for effect type main effect.

misspecifications and MSPE for G2A6 (2 active groups with 6 active variables in each) and G4A6 (4 active groups with 6 active variables). Results are similar for other settings, and are not reported for brevity. We first discuss several interesting observations for $\rho = 0$. From Figure 3, `cv.cmenet` provides a sizable improvement in selection over LASSO and SparseNet for the sibling case, but only slight improvements for the cousin and ME cases. As for predictive accuracy, the proposed method has a slight edge over the two existing methods in all three cases. The excellent performance of `cv.cmenet` in the sibling case is not unexpected, because sibling and parent-child pairs experience high correlations even in the independent setting of $\rho = 0$ (see Figure 1). By leveraging the implicit correlation structure for the CME groups, the proposed method performs much better in the presence of high collinearity. Also, when the number of active groups or variables increases, `cmenet` appears to provide improved performance, which is expected.

Next, consider the correlated setting $\rho = 1/\sqrt{2}$ in Figure 4. For all test cases, the selection performance of `cv.cmenet` is superior to both LASSO and SparseNet, providing nearly half as many misspecifications with much less variability. This is not surprising, because the CME group structure in Section 2.2 is most prominent for moderate choices of

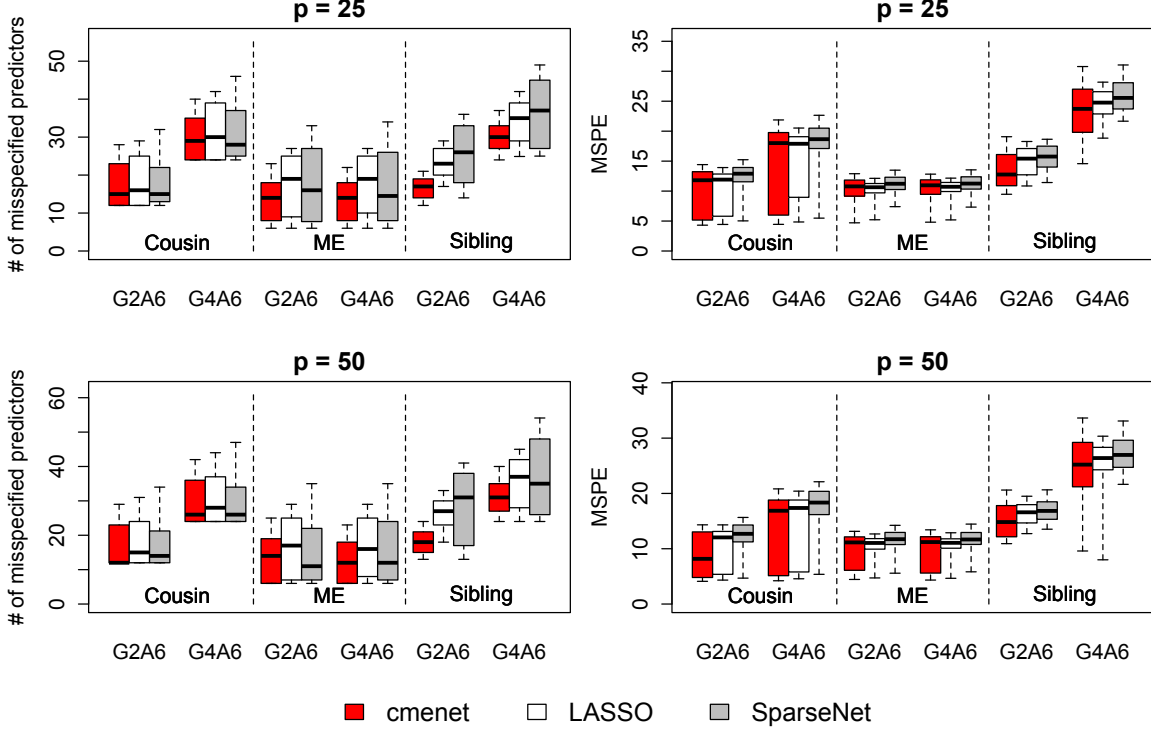


Figure 3: Boxplots of the 10%, 25%, 50%, 75% and 90% quantiles for the # of misspecified predictors (left) and MSPE (right), with the number of main effects set at $p = 25$ (top) and $p = 50$ (bottom), using a latent correlation of $\rho = 0$.

ρ . For predictive accuracy, the proposed method gives slight improvements for the cousin case, but much larger improvements for the ME and sibling cases. In light of Figure 1, this is also expected because MEs and siblings are more correlated than cousins at $\rho = 1/\sqrt{2}$.

Lastly, the tuning of selection parameters in `cv.cmenet` can also provide valuable insight on the nature of the data. For example, the average tuned value of (λ_s, λ_c) is $(6.50 \times 10^{-2}, 1.19 \times 10^{-2})$ for the cousin G2A6 test case. The fact that $\lambda_s > \lambda_c$ is quite intuitive here, because the inactive sibling effects require greater shrinkage and the active cousins require less. Likewise, the tuned values for the sibling G2A6 case is $(2.54 \times 10^{-2}, 5.25 \times 10^{-2})$, which again is expected because inactive cousins require greater shrinkage and active siblings

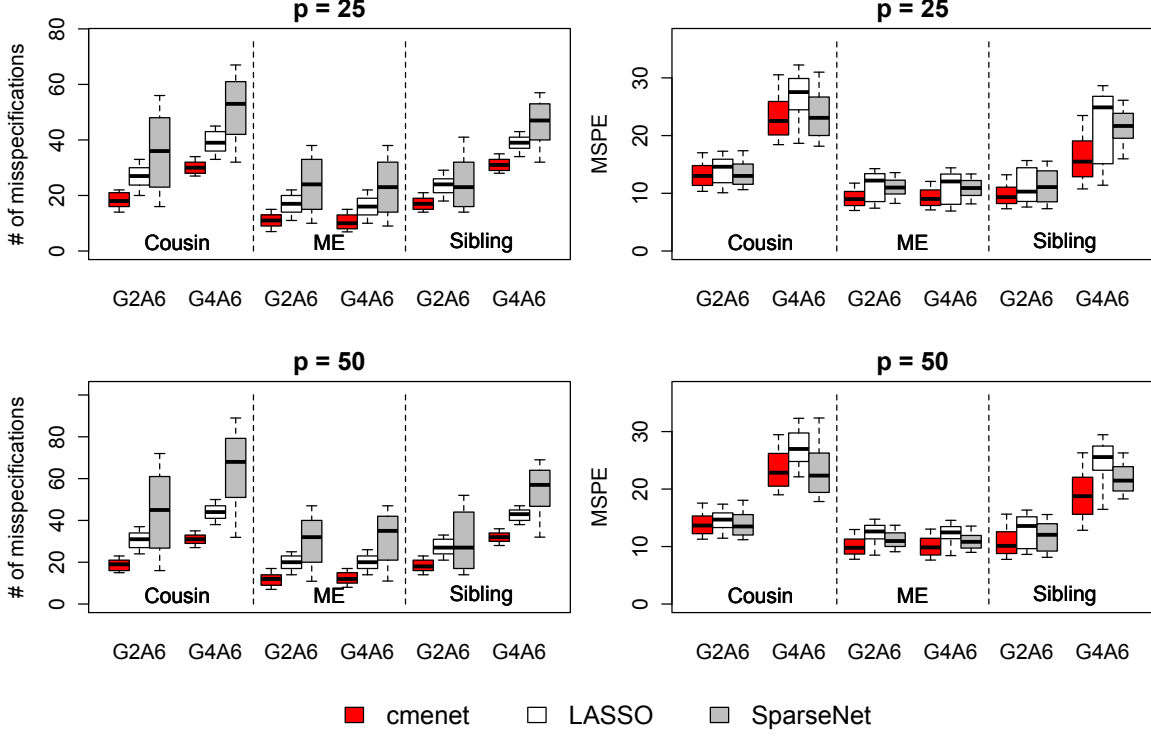


Figure 4: Boxplots of the 10%, 25%, 50%, 75% and 90% quantiles for the # of misspecified predictors (left) and MSPE (right), with the number of main effects set at $p = 25$ (top) and $p = 50$ (bottom), using a latent correlation of $\rho = 1/\sqrt{2}$.

require less. Conversely, a tuned value of $\lambda_s \approx 0$ indicates the presence of many sibling effects in data, while a tuned value of $\lambda_c \approx 0$ suggests the same for cousin effects. Given the natural interpretability of CMEs for many applications, such an analysis may reveal important insights on the data which can further guide scientific investigations.

6 Polygenic association study on fly wing shape

In this section, we demonstrate the usefulness of `cv.cmenet` for an important, real-world problem on polygenic association. Polygenes, also known as multiple gene inheritances or

quantitative genes, are a group of non-epistatic genes which affect an observable trait for an organism (Rieger et al., 2012). Such genes can serve as biological markers for many characteristics of interest, e.g., susceptibility to type 2 diabetes for youth (Rosenbloom et al., 1999) and major depressive disorders (De Moor et al., 2015). The selection of important polygenic associations from data provides an ideal application of the proposed method, because (a) predictors are binary (with +1 and -1 indicating gene presence and absence, respectively) and (b) traits with polygenic determinism are typically continuous variables (Barton and de Vlader, 2009). While such association studies are valuable tools for understanding human diseases, they can also be employed more broadly for studying genetic contributions on characteristics of other organisms. Here, we investigate the polygenic association for the wing shape of *Drosophila Melanogaster*, the common fruit fly.

The data is collected from a study by Weber et al. (2001), where the authors considered $p = 48$ polygene markers on the third chromosome of *Drosophila Melanogaster* and its effect on wing shape. In total, $n = 701$ observations are collected from recombinant isogenic lines. The response of interest is a continuous index for wing shape, which incorporates both the width of the wing across the middle and the width across the base (see Weber, 1990). The goal here is two-fold: (a) to obtain a predictive model for wing shape index, and (b) to gain further insight into the underlying genetic association from a data-driven analysis.

We compare the analysis from `cv.cmenet` with the LASSO (Tibshirani, 1997) through the R package GLMNET (Friedman et al., 2009), and SparseNet (Mazumder et al., 2011) through the R package SPARSENET (Mazumder et al., 2012). To investigate the importance of CMEs as basis functions, the latter two methods perform selection on the set of p main effects and their two-way interactions, which is the typical approach for analyzing gene-gene interactions (Cordell, 2009). This gives $p' = p + 4\binom{p}{2} = 4,560$ potential predictors for the proposed method, and $p'' = p + \binom{p}{2} = 1,176$ potential predictors for the latter two methods. In accordance with the aforementioned two-fold goal, these methods are compared on (a)

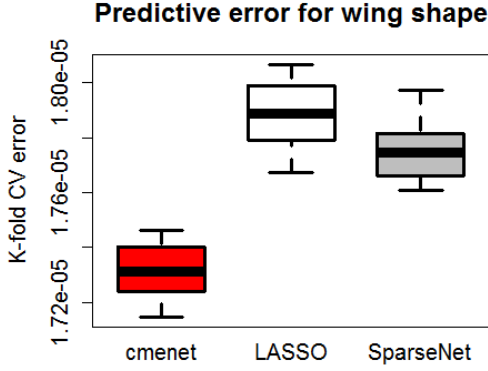


Figure 5: Boxplots of the 10%, 25%, 50%, 75% and 90% quantiles for MSPE.

Method	# of selected effects	
cv.cmenet	69	
LASSO	44	
SparseNet	35	
	Some selected effects	
cv.cmenet	V4 V1+, V4 V33+, V10 V4+, V31 V4+	V43 V1–
LASSO	V4	V43, V1 * V43
SparseNet	V4	V43, V1 * V43

Table 3: Number of effects and some selected effects for the three selection methods.

mean-squared prediction error (MSPE) and (b) the interpretability of the selected models.

Consider first Figure 5, which shows, for each of the three selection methods, the MSPE estimated using repeated 10-fold CV. `cv.cmenet` provides the best predictive model, with its highest 10% error quantile sizably smaller than the lowest 10% error quantiles for both LASSO and SparseNet. Since the key difference between the first and latter two methods is the inclusion and selection of CMEs rather than two-way interactions, such an improvement suggests the underlying polygenic association is indeed conditionally structured, i.e., certain polygenes affect wing shape *only* in the presence or absense of other polygenes. More generally, this shows that CMEs are not only appealing basis functions for studying gene association, but can also lead to greatly improved predictive models.

Next, Table 3 shows the number of selected effects for the three methods, and highlights some selected effects for each method. We see that `cv.cmenet` selected 69 active effects (MEs and CMEs), which is more than the 44 and 35 effects selected by LASSO and Sparsenet, respectively. In light of the lower predictive error for the former, this again suggests the presence of active CMEs in the data. From the first column of the selected effects, both

LASSO and SparseNet deemed effect $V4$ (i.e., the fourth polygene) to be active, while `cv.cmenet` instead selected the two sibling effects $V4|V1+$, $V4|V33+$ and the two cousin effects $V10|V4+$, $V31|V4+$. Put another way, under the two existing models, the gene $V4$ is considered influential for wing shape under all situations, whereas under the new model, the conclusion is more nuanced, with $V4$ influential only when genes $V1$ or $V33$ are active, or in activating genes $V10$ or $V31$. The latter provides a more careful analysis of the signal from $V4$, and judging from MSPE, more accurately identifies the underlying gene association structure. Similarly, from the second column of the selected effects, both $V43$ and $V1 * V43$ are active under the two existing methods, while only $V43|V1-$ is active for `cv.cmenet`. The substitution of $V43$ and $V1 * V43$ for $V43|V1-$ is precisely the first rule in Su and Wu (2017) for CME selection in designed experiments, where an active ME and its interaction is replaced with its corresponding CME for a more parsimonious model.

To summarize, this gene association study highlights two advantages of `cmenet`. First, in applications where CMEs are interpretable phenomena, including them as basis vectors for selection can greatly reduce prediction error. Second, compared to an analysis of two-way interactions, which can be difficult to interpret, the proposed selection method can provide greater insight on the underlying problem, and can guide further scientific investigations. This is particularly true for genetic applications, where selected CMEs can be used to further investigate why some genes are *conditionally* active, and why some play a more supportive role in *activating* other genes.

7 Conclusion and future work

In this paper, a new method is presented for selecting binary predictors and a set of reparametrized predictors called conditional main effects (CMEs) from observation data. While CMEs are intuitive basis functions with appealing interpretations in many applications,

existing selection methods perform poorly due to the inherent grouped collinearity structure. We proposed a novel selection method called **cmenet**, which accounts for CME groupings using sibling, cousin, parent-child and uncle-nephew pairs. **cmenet** offers two attractive features called CME coupling and reduction, with the former allowing CMEs to more easily enter the model given selected siblings or cousins, and the latter encouraging the selection of the underlying ME given many selected siblings or cousins. A coordinate descent algorithm is introduced for minimizing the selection criterion, and several computational tools are then proposed for efficient optimization and parameter tuning in high-dimensions. Simulation studies showed sizable improvements for **cmenet** over existing methods with respect to selection and prediction accuracy. Applying this to a real-world gene association dataset, the proposed method provides not only reduced predictive error, but also a highly interpretable model which reveals important insights on gene activation behavior.

Given the positive results here, there are many exciting avenues for future work. First, in the high dimensional setting of $p \gg 1$, the tuning of four selection parameters in $Q(\beta)$ can be quite expensive computationally. One reason for this is the use of a grid structure for testing feasible parameter combinations in **cv.cmenet**. With recent advances on the topic of optimal designs for convex spaces (e.g., Lekivetz and Jones, 2015; Mak and Joseph, 2016), it may be interesting to see whether the use of such designs as candidate settings allows for more efficient parameter tuning. Second, we are working to broaden the proposed methodology to higher-order conditional effects, e.g., the effect of A conditional on both $B+$ and $C+$. The main challenge here is again computational efficiency, but such a direction would enable the investigation of, say, more complex activation phenomena in the earlier gene study. Lastly, we are also interested in extending CMEs for quantifying the conditional effects of *continuous* and *multi-level* factors, because such an extension allows the proposed methodology to be applicable for more general datasets.

An efficient C++ implementation of **cmenet** and **cv.cmenet** is available in the R package

CMENET in CRAN. The authors gratefully acknowledge helpful advice from Prof. Ben Haaland.

References

- Barton, N. H. and de Vladar, H. P. (2009). Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics*, 181(3):997–1011.
- Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.
- Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369.
- Chari, S. and Dworkin, I. (2013). The conditional nature of genetic interactions: the consequences of wild-type backgrounds on mutational interactions in a genome-wide modifier screen. *PLoS Genetics*, 9(8):e1003661.
- Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404.
- De Moor, M. H., Van Den Berg, S. M., Verweij, K. J., Krueger, R. F., Luciano, M., Vasquez, A. A., Matteson, L. K., Derringer, J., Esko, T., Amin, N., et al. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry*, 72(7):642–650.
- Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- Finney, D. (1945). The fractional replication of factorial arrangements. *Annals of Eugenics*, 12:291–303.

- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.
- Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer.
- Friedman, J., Hastie, T., and Tibshirani, R. (2009). GLMNET: Lasso and elastic-net regularized generalized linear models. *R package version 1*.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.
- Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.
- Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440.
- Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer Science & Business Media.
- Lekivetz, R. and Jones, B. (2015). Fast flexible space-filling designs for nonrectangular regions. *Quality and Reliability Engineering International*, 31(5):829–837.
- Leung, G. P., Aristizabal, M. J., Krogan, N. J., and Kobor, M. S. (2014). Conditional genetic interactions of RTT107, SLX4, and HRQ1 reveal dynamic networks upon dna damage in *S. cerevisiae*. *G3: Genes—Genomes—Genetics*, 4(6):1059–1069.
- Mak, S. and Joseph, V. R. (2016). Minimax and minimax projection designs using clustering. *arXiv preprint arXiv:1602.03938*.
- Mazumder, R., Friedman, J. H., and Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495).
- Mazumder, R., Hastie, T., and Friedman, J. (2012). SPARSENET: Fit sparse linear regression

- models via nonconvex optimization. *R package version 1*.
- Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71.
- Montgomery, D. C. (2008). *Design and Analysis of Experiments*. John Wiley & Sons.
- Rieger, R., Michaelis, A., and Green, M. M. (2012). *Glossary of Genetics: Classical and Molecular*. Springer Science & Business Media.
- Rosenbloom, A. L., Joe, J. R., Young, R. S., and Winter, W. E. (1999). Emerging epidemic of type 2 diabetes in youth. *Diabetes Care*, 22(2):345–354.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.
- Stuart, A. and Ord, J. (1994). *Kendall’s Advanced Theory of Statistics, Volume 1: Distribution Theory*. Arnold London.
- Su, H. and Wu, C. F. J. (2017). CME analysis: a new method for unraveling aliased effects in two-level fractional factorial experiments. *Journal of Quality Technology*, to appear.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B*, 74(2):245–266.
- Weber, K. (1990). Selection on wing allometry in *Drosophila Melanogaster*. *Genetics*, 126(4):975–989.
- Weber, K., Eisman, R., Higgins, S., Morey, L., Patty, A., Tausek, M., and Zeng, Z.-B. (2001). An analysis of polygenes affecting wing shape on chromosome 2 in *Drosophila Melanogaster*. *Genetics*, 159(3):1045–1057.

- Wu, C. F. J. (2015). Post-Fisherian experimentation: from physical to virtual. *Journal of the American Statistical Association*, 110(510):612–620.
- Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, volume 552. John Wiley & Sons.
- Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Appendix

Proof of Theorem 1:

The proof of this requires a simple lemma on normal orthant probabilities:

Lemma 5. (*Stuart and Ord, 1994*) Let (X_1, \dots, X_p) follow the equicorrelated normal distribution, with $\mathbb{E}(X_j) = 0$, $\mathbb{E}(X_j^2) = 1$ and $\mathbb{E}(X_j X_k) = \rho$ for all $j \neq k$, and let $p_m = \mathbb{P}(X_1 > 0, \dots, X_m > 0)$. Then:

$$p_2 = \frac{\sin^{-1} \rho}{2\pi} + \frac{1}{4} \quad \text{and} \quad p_3 = \frac{3 \sin^{-1} \rho}{4\pi} + \frac{1}{8}.$$

For the main proof, note that each row of the latent matrix \mathbf{Z} is i.i.d., so it suffices to fix $n = 1$ and explore the correlation amongst the scalar ME quantities $\tilde{x}_{1,A}$ and CME quantities $\tilde{x}_{1,A|B+}$. We denote these as \tilde{x}_A and $\tilde{x}_{A|B+}$ for brevity. Under the latent equicorrelated distribution $\mathcal{N}\{\mathbf{0}, \rho\mathbf{J} + (1 - \rho)\mathbf{I}\}$, it is easy to show that $\mathbb{E}[\tilde{x}_A] = 0$ and $\text{Var}[\tilde{x}_A] = 1$. Moreover, the CME $\tilde{x}_{A|B+}$ can be conditionally decomposed as $\tilde{x}_{A|B+} \stackrel{d}{=} R[2p_2]$ if $\tilde{x}_B = +1$, and 0 if $\tilde{x}_B = -1$, where $R[q]$ is the Rademacher random variable taking on $+1$ w.p. $q \in [0, 1]$ and -1 otherwise. From this, we get:

$$\begin{aligned}\mu_c &\equiv \mathbb{E}[\tilde{x}_{A|B+}] = \mathbb{E}[\mathbb{E}[\tilde{x}_{A|B+}|\tilde{x}_B]] = \frac{1}{2}(4p_2 - 1), \\ \sigma_c^2 &\equiv \text{Var}[\tilde{x}_{A|B+}] = \text{Var}[\mathbb{E}[\tilde{x}_{A|B+}|\tilde{x}_B]] + \mathbb{E}[\text{Var}[\tilde{x}_{A|B+}|\tilde{x}_B]] = \frac{1}{2} - \left(\frac{\sin^{-1}\rho}{\pi}\right)^2.\end{aligned}$$

Consider the correlation between the MEs \tilde{x}_A and \tilde{x}_B . Note that $\tilde{x}_A\tilde{x}_B$ equals $+1$ when \tilde{x}_A and \tilde{x}_B have the same sign, and equals -1 otherwise. Letting $\mathbb{P}(++)$ be the probability of $(\tilde{x}_A, \tilde{x}_B) = (+1, +1)$ (with similar notation for $+-$, $-+$ and $--$), Lemma 5 then gives:

$$\text{Corr}(\tilde{x}_A, \tilde{x}_B) = [\mathbb{P}(++) + \mathbb{P}(--)] - [\mathbb{P}(+-) + \mathbb{P}(-+)] = 2p_2 - 2[1/2 - p_2] = \frac{2\sin^{-1}\rho}{\pi}.$$

Next, consider the two sibling CMEs $\tilde{x}_{A|B+}$ and $\tilde{x}_{A|C+}$. Note that $\tilde{x}_{A|B+}\tilde{x}_{A|C+}$ equals $+1$ when both $\tilde{x}_B = +1$ and $\tilde{x}_C = +1$, and equals 0 otherwise. It follows that:

$$\text{Corr}(\tilde{x}_{A|B+}, \tilde{x}_{A|C+}) = \frac{1}{\sigma_c^2}[\mathbb{P}(++) - \mu_c^2] = \frac{1}{\sigma_c^2}[p_2 - \mu_c^2] = \frac{1}{\sigma_c^2} \left\{ -\left(\frac{\sin^{-1}\rho}{\pi}\right)^2 + \frac{\sin^{-1}\rho}{2\pi} + \frac{1}{4} \right\}.$$

Consider now the two cousin CMEs $\tilde{x}_{B|A+}$ and $\tilde{x}_{C|A+}$. Note that $\tilde{x}_{B|A+}\tilde{x}_{C|A+}$ equals $+1$ when $\tilde{x}_A = +1$ and $\tilde{x}_B = \tilde{x}_C$, $\tilde{x}_{B|A+}\tilde{x}_{C|A+}$ equals -1 when $\tilde{x}_A = +1$ and $\tilde{x}_B \neq \tilde{x}_C$, and equals

0 otherwise. We then have:

$$\begin{aligned}
\text{Corr}(\tilde{x}_{B|A+}, \tilde{x}_{C|A+}) &= \frac{1}{\sigma_c^2} [\{\mathbb{P}(+++)+\mathbb{P}(+--)\}-\{\mathbb{P}(++-)+\mathbb{P}(+-+)\}-\mu_c^2] \\
&= \frac{1}{\sigma_c^2} [\{\mathbb{P}(+++)+(\mathbb{P}(--)-\mathbb{P}(---))\}-2\{\mathbb{P}(++)-\mathbb{P}(+++)\}-\mu_c^2] \\
&= \frac{1}{\sigma_c^2} [2p_3 - p_2 - \mu_c^2] = \frac{1}{\sigma_c^2} \left\{ - \left(\frac{\sin^{-1} \rho}{\pi} \right)^2 + \frac{\sin^{-1} \rho}{\pi} \right\}.
\end{aligned}$$

Correlations for the remaining two effect pairs can be proved in a similar manner.

Proof of Theorem 2:

Let $\mathbf{X} \in \mathbb{R}^{n \times p'}$ be the normalized model matrix consisting of all main effects and CMEs, where $p' = p + 4\binom{p}{2}$. By the strong law of large numbers, the sample covariance matrix $\mathbf{C}_n = \mathbf{X}^T \mathbf{X} / n$ converges elementwise to some matrix $\mathbf{C} \in \mathbb{R}^{p' \times p'}$ with unit diagonal entries and off-diagonal entries given in Theorem 1. Consider the following block partition of $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$, where \mathbf{C}_{11} is the block for the active set \mathcal{A} , and \mathbf{C}_{22} the block for the remaining variables. Zhao and Yu (2006) proved that the LASSO is sign-selection consistent only when the (weak) *irrepresentability condition* holds: $\forall \boldsymbol{\zeta} \in \{-1, +1\}^{p'}, |\mathbf{C}_{21} \mathbf{C}_{11}^{-1} \boldsymbol{\zeta}| < \mathbf{1}$ (this is a slight simplification of the original condition under the current i.i.d. setting). Hence, sign-selection inconsistency can be proven if $\exists \boldsymbol{\zeta} \in \{-1, +1\}^{p'}$ and an inactive effect j satisfying:

$$|\mathbf{C}_{21,j} \mathbf{C}_{11}^{-1} \boldsymbol{\zeta}| \geq 1, \quad \text{where } \mathbf{C}_{21,j} \text{ is the row corresponding to effect } j. \quad (9)$$

Consider first a model with only $q \geq 3$ active siblings of the form $A|B+$, $A|C-$, ..., $A|R-$. Using the same principles as in Theorem 1, \mathbf{C}_{11} can be shown to be a $q \times q$ matrix

with unit diagonal, $[(1/2 - p_2) - \mu_c^2]/\sigma_c^2$ for off-diagonal entries in the first row and column, and $\psi_{sib}(\rho)$ for all other off-diagonal entries. Letting A be the inactive effect, we have $\mathbf{C}_{21,A} = \psi_{pc}(\rho)\mathbf{1}_q^T$, and letting $\boldsymbol{\zeta} = \mathbf{1}_q$, it follows that $|\mathbf{C}_{21,A}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0$. By (9), part (a) is proven.

Next, consider a model with only $q = 2$ active main effects, say, A and $-B$. From Theorem 1, \mathbf{C}_{11} is a $q \times q$ matrix with unit diagonal and $-\psi_{me}(\rho)$ on the off-diagonals. Let $A|B-$ be the inactive effect, so $\mathbf{C}_{21,A|B-} = (\psi_{pc}(\rho), \psi_{un}(\rho))$. Taking $\boldsymbol{\zeta} = (1, 1)^T$, $|\mathbf{C}_{21,A|B-}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0.27$, thereby proving selection inconsistency.

Lastly, consider a model with only $q \geq 6$ active cousins of the form $B|A+$, $C|A-$, ..., $R|A-$. Using the same principles as in Theorem 1, \mathbf{C}_{11} is a $q \times q$ matrix with unit diagonal, $-\mu_c^2/\sigma_c^2$ for the off-diagonal entries in the first row and column, and $\psi_{cou}(\rho)$ for all other off-diagonal entries. Let B be the inactive effect with $\mathbf{C}_{21,B} = (\psi_{sib}(\rho), \psi_{un}(\rho)\mathbf{1}_{q-1})$. Taking $\boldsymbol{\zeta} = \mathbf{1}_q$, $|\mathbf{C}_{21,B}\mathbf{C}_{11}^{-1}\boldsymbol{\zeta}| \geq 1$ for $\rho \geq 0.29$, which proves inconsistency.

Proof of Proposition 1:

As a note, since the objective $Q(\boldsymbol{\beta})$ is non-differentiable at $\boldsymbol{\beta} = \mathbf{0}$, what we mean by strict convexity here is that $\nabla_{\mathbf{u}}^2 Q(\boldsymbol{\beta})$, the directional Hessian of $Q(\boldsymbol{\beta})$ in direction \mathbf{u} , is strictly positive for all $\boldsymbol{\beta}$ and all $\|\mathbf{u}\| = 1$. We follow a similar approach as Proposition 1 of Breheny (2015). Note that $\nabla^2 \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 2\mathbf{X}^T\mathbf{X}$. Moreover, because $\eta''_{\lambda,\tau}(\theta) = -\tau \exp(-\theta\tau/\lambda)$, it follows that $\nabla_{\mathbf{u}}^2 P_s(\boldsymbol{\beta}) \geq -\tau$ and $\nabla_{\mathbf{u}}^2 P_c(\boldsymbol{\beta}) \geq -\tau$ for all \mathbf{u} and $\boldsymbol{\beta}$. Hence:

$$\nabla_{\mathbf{u}}^2 Q(\boldsymbol{\beta}) = \nabla_{\mathbf{u}}^2 \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_s(\boldsymbol{\beta}) + P_c(\boldsymbol{\beta}) \right\} \geq \frac{\lambda_{\min}(\mathbf{X}^T\mathbf{X})}{n} - 2\tau \text{ for all } \mathbf{u} \text{ and } \boldsymbol{\beta},$$

which is strictly positive when $\tau < \lambda_{\min}(\mathbf{X}^T\mathbf{X})/(2n)$. The second part of the claim follows by replacing \mathbf{X} with \mathbf{x}_j in the argument above, and using the fact that $\|\mathbf{x}_j\|_2^2 = 1$.

Proof of Theorem 3 and Corollary 1:

The majorization claim *a*) follows from a first-order Taylor expansion of the outer penalty: $\eta_{\lambda,\tau}(\|\boldsymbol{\beta}_g\|_{\lambda,\gamma}) \geq \eta_{\lambda,\tau}(\|\tilde{\boldsymbol{\beta}}_g\|_{\lambda,\gamma}) + \tilde{\Delta}_g \left\{ \|\boldsymbol{\beta}_g\|_{\lambda,\gamma} - \|\tilde{\boldsymbol{\beta}}_g\|_{\lambda,\gamma} \right\}$, where the inequality is a result of the concavity of η . See Lemma 1 in Breheny (2015) for details.

For the threshold function in *b*), take the following optimization problem:

$$\hat{\beta}_j = \operatorname{argmin}_{\beta_j} \left\{ \frac{1}{2n} \|\mathbf{r} - \mathbf{x}_j \beta_j\|_2^2 + \Delta_1 g_{\lambda_1, \gamma}(\beta_j) + \Delta_2 g_{\lambda_2, \gamma}(\beta_j) \right\}. \quad (10)$$

The KKT condition for (10) is:

$$0 \in -\frac{1}{n} \mathbf{x}_j^T \mathbf{r} + \hat{\beta}_j + \Delta_1 \partial_{\lambda_1, \gamma} \hat{\beta}_j + \Delta_2 \partial_{\lambda_2, \gamma} \hat{\beta}_j, \quad \partial_{\lambda, \gamma} \beta_j = \begin{cases} \operatorname{sgn}(\beta_j) \left(1 - \frac{|\beta_j|}{\lambda \gamma} \right)_+ & \text{if } |\beta_j| > 0, \\ [-1, 1] & \text{if } \beta_j = 0. \end{cases} \quad (11)$$

Without loss of generality, assume $z \equiv \mathbf{x}_j^T \mathbf{r}/n > 0$. Consider the same four cases for z as presented in (8):

1. $z \geq \lambda_{(1)}\gamma$: Suppose $\hat{\beta}_j = z$. Then the KKT condition (11) becomes $0 \in -z + \hat{\beta}_j$, which is satisfied. Since (10) is strictly convex, $\hat{\beta}_j = z$ must be its unique solution.
2. $c_2 \leq z < \lambda_{(1)}\gamma$ (see (8) for c_2): Suppose $\hat{\beta}_j = (z - \Delta_{(1)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma} \right)$. Since $\lambda_{(2)}\gamma \leq \hat{\beta}_j < \lambda_{(1)}\gamma$, the KKT condition (11) becomes $0 \in -z + \hat{\beta}_j + \Delta_{(1)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(1)}\gamma} \right)$, which is satisfied. Hence, $\hat{\beta}_j$ is the unique solution to (10).
3. $\Delta_{(1)} + \Delta_{(2)} \leq z < c_2$ (see (8) for c_3): Suppose $\hat{\beta}_j = (z - \Delta_{(1)} - \Delta_{(2)}) / \left(1 - \frac{\Delta_{(1)}}{\lambda_{(1)}\gamma} - \frac{\Delta_{(2)}}{\lambda_{(2)}\gamma} \right)$. Since $0 < \hat{\beta}_j < \lambda_{(2)}\gamma$, the KKT condition (11) becomes $0 \in -z + \hat{\beta}_j + \Delta_{(1)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(1)}\gamma} \right) + \Delta_{(2)} \left(1 - \frac{\hat{\beta}_j}{\lambda_{(2)}\gamma} \right)$, which is satisfied. Hence, $\hat{\beta}_j$ is the unique solution to (10).
4. $0 \leq z < \Delta_{(1)} + \Delta_{(2)}$: Suppose $\hat{\beta}_j = 0$. The KKT condition then becomes $0 \in -z + (\Delta_{(1)} + \Delta_{(2)})[-1, 1]$, which is satisfied, so $\hat{\beta}_j$ is the unique solution to (10).

From this, Corollary 1 can be proved in a similar way as Proposition 3 of Breheny (2015).

Proof of Proposition 2:

Since $Q(\boldsymbol{\beta})$ is strictly convex, it must have at most one minimizer $\boldsymbol{\beta}$. By definition, $\boldsymbol{\beta}$ must satisfy the KKT condition:

$$0 \in -\frac{1}{n}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \Delta_S(\boldsymbol{\beta})\partial_{\lambda_s, \gamma}\beta_j + \Delta_C(\boldsymbol{\beta})\partial_{\lambda_c, \gamma}\beta_j, \quad j = 1, \dots, p', \quad (12)$$

where $\partial_{\lambda, \gamma}\beta_j$ is the subgradient defined in (11), and $\Delta_S(\boldsymbol{\beta})$ and $\Delta_C(\boldsymbol{\beta})$ are the linearized slopes in (4) for the sibling and cousin groups of effect j . Setting $\boldsymbol{\beta} = \mathbf{0}$, (12) becomes:

$$-\frac{1}{n}\mathbf{x}_j^T\mathbf{y} + \lambda_s[-1, 1] + \lambda_c[-1, 1] = -\frac{1}{n}\mathbf{x}_j^T\mathbf{y} + [-\lambda_s - \lambda_c, \lambda_s + \lambda_c],$$

which contains 0 when $\lambda_s + \lambda_c \geq |\mathbf{x}_j^T\mathbf{y}|$. Hence, when $\lambda + \lambda_c \geq \max_{j=1, \dots, p'} |\mathbf{x}_j^T\mathbf{y}|$, the trivial solution $\boldsymbol{\beta} = \mathbf{0}$ is the unique minimizer of $Q(\boldsymbol{\beta})$.

Proof of Theorem 4:

Unless necessary, we refer to the optimal coefficient $\hat{\beta}_j(\lambda_s^l)$ as $\hat{\beta}_j^l$ to avoid unnecessary notation. For a given coefficient solution $\boldsymbol{\beta}$ with $0 \leq \beta_j < \gamma \min(\lambda_s^l, \lambda_c)$, the KKT condition for effect j under penalty λ_s^l becomes:

$$c_j(\lambda_s^l) \in \Delta_S(\boldsymbol{\beta})\partial_{\lambda_s^l, \gamma}\beta_j + \Delta_C(\boldsymbol{\beta})\partial_{\lambda_c, \gamma}\beta_j = \begin{cases} \left\{ \text{sgn}(\beta_j)D(\beta_j; \lambda_s^l) (1 - |\beta_j|(\lambda_s^l\gamma)^{-1}) \right. \\ \quad \left. + \text{sgn}(\beta_j)D(\beta_j; \lambda_c) (1 - |\beta_j|(\lambda_c\gamma)^{-1}) \right\}, & \text{if } \beta_j \neq 0 \\ D(\beta_j; \lambda_s^l) \cdot [-1, 1] + D(\beta_j; \lambda_c) \cdot [-1, 1], & \text{if } \beta_j = 0. \end{cases} \quad (13)$$

Here, $D(\beta_j; \lambda) = \lambda \exp\{-(\tau/\lambda)g_{\lambda, \gamma}(\beta_j)\}$ is the linearized slope under Assumption A.3,

where only effect j enters the model without other effects in its sibling or cousin group.

Now, suppose $c_j(\lambda_s^{l+1})$ is known in advance. By assumption, $\hat{\beta}_j^{l+1} = 0$, i.e., effect j is inactive under the previous penalty λ_s^{l+1} . Assuming $\hat{\beta}_j(\lambda_s)$ is $\overline{\nabla\beta}$ -Lipschitz in λ_s (this is proved later in Lemma 6), we have $|\hat{\beta}_j^{l+1} - \hat{\beta}_j^l| \leq \overline{\nabla\beta}(\lambda_s^{l+1} - \lambda_s^l) = \overline{\Delta\beta_l}$, so the largest possible value of $\hat{\beta}_j^l$ is $\overline{\Delta\beta_l}$. Using this upper bound along with the KKT condition (13), a sufficient condition for $\hat{\beta}_j^l = 0$ is:

$$|c_j(\lambda_s^l)| < D(\overline{\Delta\beta_l}; \lambda_s^l) \left(1 - \frac{\overline{\Delta\beta_l}}{\lambda_s^l \gamma}\right) + D(\overline{\Delta\beta_l}; \lambda_c) \left(1 - \frac{\overline{\Delta\beta_l}}{\lambda_c \gamma}\right), \quad (14)$$

since by Assumption A.2, $\overline{\Delta\beta_l} \leq \gamma \min(\lambda_s^l, \lambda_c)$.

Consider next the following decomposition $|c_j(\lambda_s^l)| \leq |c_j(\lambda_s^{l+1}) - c_j(\lambda_s^l)| + |c_j(\lambda_s^l)|$. By Assumption A.1, $|c_j(\lambda_s^{l+1}) - c_j(\lambda_s^l)|$ must be bounded above by $(\lambda_s^{l+1} - \lambda_s^l)/n$. The second term $|c_j(\lambda_s^l)|$ must also be bounded by M_l (see definition in theorem). Putting these together, we get:

$$|c_j(\lambda_s^l)| < |\lambda_s^{l+1} - \lambda_s^l|/n + M_l = D(\overline{\Delta\beta_l}; \lambda_s^l) \left(1 - \frac{\overline{\Delta\beta_l}}{\lambda_s^l \gamma}\right) + D(\overline{\Delta\beta_l}; \lambda_c) \left(1 - \frac{\overline{\Delta\beta_l}}{\lambda_c \gamma}\right),$$

which is precisely the condition in (14). Hence, $\hat{\beta}_j^l = 0$, as desired.

All that is left is to prove the Lipschitz condition on $\hat{\beta}_j(\lambda_s)$. Under the mild assumption that $\hat{\beta}_j(\lambda_s)$ is continuous and almost-everywhere differentiable (not proven here), this condition is equivalent to $|\frac{\partial}{\partial \lambda_s} \hat{\beta}_j(\lambda_s)|$ being upper bounded by $\overline{\nabla\beta}$. The latter is proved in the lemma below:

Lemma 6. $\left| \frac{\partial}{\partial \lambda_s} \hat{\beta}_j(\lambda_s) \right| \leq \overline{\nabla\beta}$ whenever $0 < \hat{\beta}_j(\lambda_s) < \gamma \min(\lambda_s, \lambda_c)$.

Proof. For $0 < \hat{\beta}_j(\lambda_s) < \gamma \min(\lambda_s, \lambda_c)$, the KKT condition becomes:

$$D(\hat{\beta}_j; \lambda_s) \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right) + D(\hat{\beta}_j; \lambda_c) \left(1 - \frac{\hat{\beta}_j}{\lambda_c \gamma}\right) = c_j(\lambda_s).$$

Taking first the derivative with respect to λ_s , then the absolute value on both sides, we get:

$$\left| \frac{\partial}{\partial \lambda_s} \left\{ D(\hat{\beta}_j; \lambda_s) \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right) + D(\hat{\beta}_j; \lambda_c) \left(1 - \frac{\hat{\beta}_j}{\lambda_c \gamma}\right) \right\} \right| = \left| \frac{\partial}{\partial \lambda_s} c_j(\lambda_s) \right| \leq \frac{1}{n},$$

where the inequality holds from the contractive condition in Assumption A.1.

Using the chain rule, the Liebniz integral rule, and simple (but tedious) algebra, the above expression becomes:

$$\begin{aligned} & \left| \left[\frac{D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right) \left\{ 1 - \frac{\tau}{\lambda_s} \left(\frac{\hat{\beta}_j}{\lambda_s \gamma} - \hat{\beta}_j \right) \right\} + \frac{D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \frac{\hat{\beta}_j}{\lambda_s \gamma} \right] \right. \\ & \left. - \left[\frac{\tau D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right)^2 + \frac{D(\hat{\beta}_j; \lambda_s)}{\gamma \lambda_s} + \frac{\tau D(\hat{\beta}_j; \lambda_c)}{\lambda_c} \left(1 - \frac{\hat{\beta}_j}{\lambda_c \gamma}\right)^2 + \frac{D(\hat{\beta}_j; \lambda_c)}{\gamma \lambda_c} \right] \frac{\partial}{\partial \lambda_s} \hat{\beta}_j \right| \leq \frac{1}{n}. \end{aligned}$$

Finally, since $||x| - |y|| \leq |x - y|$ for $x, y \in \mathbb{R}$, it follows that:

$$\begin{aligned} \left| \frac{\partial}{\partial \lambda_s} \hat{\beta}_j \right| & \leq \frac{\frac{1}{n} + \frac{D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right) \left\{ 1 - \frac{\tau}{\lambda_s} \left(\frac{\hat{\beta}_j}{\lambda_s \gamma} - \hat{\beta}_j \right) \right\} + \frac{D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \frac{\hat{\beta}_j}{\lambda_s \gamma}}{\frac{\tau D(\hat{\beta}_j; \lambda_s)}{\lambda_s} \left(1 - \frac{\hat{\beta}_j}{\lambda_s \gamma}\right)^2 + \frac{D(\hat{\beta}_j; \lambda_s)}{\gamma \lambda_s} + \frac{\tau D(\hat{\beta}_j; \lambda_c)}{\lambda_c} \left(1 - \frac{\hat{\beta}_j}{\lambda_c \gamma}\right)^2 + \frac{D(\hat{\beta}_j; \lambda_c)}{\gamma \lambda_c}} \\ & \leq \frac{\frac{1}{n} + \left(1 + \frac{\tau \gamma}{2}\right) + 1}{\frac{D(\gamma \lambda_s; \lambda_s)}{\gamma \lambda_s} + \frac{D(\gamma \lambda_c; \lambda_c)}{\gamma \lambda_c}} = \frac{\frac{1}{n} + 2 + \frac{\tau \gamma}{2}}{\frac{2}{\gamma} \exp\left\{-\frac{\tau \gamma}{2}\right\}} = \overline{\nabla \beta}, \end{aligned}$$

where the last inequality holds because $0 < \hat{\beta}_j < \gamma \min(\lambda_s, \lambda_c)$. □

Algorithm statement for `cv.cmenet`:

Algorithm 2 `cv.cmenet`: a cross-validation algorithm for tuning `cmenet`

```

1: function CV.CMENET( $\mathbf{X}, \mathbf{y}$ )
2:   • Initialize  $0 < \lambda_1 < \dots < \lambda_L < \max_{j=1, \dots, p'} |\mathbf{x}_j^T \mathbf{y}|$ ,  $0 < \gamma_1 < \dots < \gamma_G$  and  $0 < \tau_1 < \dots < \tau_T < 1/(2n)$ .
3:   • Tune  $(\lambda^*, \gamma^*)$  using cv.sparsenet in the R package SPARSENET, and set  $\lambda_s^*, \lambda_c^* \leftarrow \lambda^*/2$ .
4:   • Randomly partition the data  $\mathcal{D} = (\mathbf{X}, \mathbf{y})$  into  $K$  equal pieces  $\{\mathcal{D}_1, \dots, \mathcal{D}_K\}$ .
5:   for  $k = 1, \dots, K$  do                                     ▷  $K$ -fold CV for tuning  $\gamma$  and  $\tau$ 
6:     for  $\gamma \in \{\gamma_1, \dots, \gamma_G\}$  do                               ▷ For each  $\gamma \dots$ 
7:       •  $\beta_{prev} \leftarrow \mathbf{0}_{p'}$ .                                     ▷ Reset warm start solution
8:       for  $\tau \in \{\tau_1, \dots, \tau_T\}$  do                               ▷ For each  $\tau \dots$ 
9:         •  $\beta_{\lambda_s^*, \lambda_c^*}(\gamma, \tau; k) \leftarrow \text{cmenet}(\mathbf{X}_{-k}, \mathbf{y}_{-k}, \lambda_s^*, \lambda_c^*, \gamma, \tau, \beta_{prev})$  ▷ Train w/o part  $k$ 
10:        •  $\beta_{prev} \leftarrow \beta_{\lambda_s^*, \lambda_c^*}(\gamma, \tau; k)$                 ▷ Update warm start solution
11:   •  $(\gamma^*, \tau^*) \leftarrow \underset{\gamma, \tau}{\operatorname{argmin}} \sum_{k=1}^K \|\mathbf{y}_k - \mathbf{X}_k \beta_{\lambda_s^*, \lambda_c^*}(\gamma, \tau; k)\|_2$  ▷ Estimate optimal  $\gamma$  and  $\tau$ 
12:   for  $k = 1, \dots, K$  do                                     ▷  $K$ -fold CV for tuning  $\lambda_s$  and  $\lambda_c$ 
13:     for  $\lambda_c \in \{\lambda_L, \dots, \lambda_1\}$  do                               ▷ For each  $\lambda_c \dots$ 
14:       •  $\beta_{prev} \leftarrow \mathbf{0}_{p'}$ .
15:       for  $\lambda_s \in \{\lambda_L, \dots, \lambda_1\}$  do                               ▷ For each  $\lambda_s \dots$ 
16:         if  $\lambda_c + \lambda_s < \max_{j=1, \dots, p'} |\mathbf{x}_j^T \mathbf{y}|$  then
17:           •  $\beta_{\lambda_s, \lambda_c}(\gamma^*, \tau^*; k) \leftarrow \text{cmenet}(\mathbf{X}_{-k}, \mathbf{y}_{-k}, \lambda_s, \lambda_c, \gamma^*, \tau^*, \beta_{prev})$ , using  $\beta_{prev}$  and Theorem 4 for screening.
18:           •  $\beta_{prev} \leftarrow \beta_{\lambda_s, \lambda_c}(\gamma^*, \tau^*; k)$ 
19:   •  $(\lambda_s^*, \lambda_c^*) \leftarrow \underset{\lambda_s, \lambda_c}{\operatorname{argmin}} \sum_{k=1}^K \|\mathbf{y}_k - \mathbf{X}_k \beta_{\lambda_s, \lambda_c}(\gamma^*, \tau^*; k)\|_2$  ▷ Estimate optimal  $\lambda_s$  and  $\lambda_c$ 
20:   •  $\hat{\beta} \leftarrow \text{cmenet}(\mathbf{X}, \mathbf{y}, \lambda_s^*, \lambda_c^*, \gamma^*, \tau^*, \mathbf{0}_{p'})$  ▷ Refit using optimal parameters
   return optimal coefficients  $\hat{\beta}$ .

```
